

10G Ethernet: Now Ready for Low-Latency HPC Applications

Solarflare extends the benefits of its low-latency, high-bandwidth 10GbE server adapters to High-Performance Computing (HPC) applications



INTRODUCTION

High-performance computing clusters rely on predictable low-latency networking to tune inter-process communications, as well as sufficient network bandwidth to support scalable clusters, storage access and management. HPC clusters require high capacity, low power, non-blocking fabrics that are easy to deploy, easy to manage, and cost-effective. Their deployment requires a flat switching architecture implemented with high-speed top of rack switches with low-latency network adapters to transfer large amounts of data between huge server memories in each compute node.

Until now HPC deployments generally had to choose between high performance (Infiniband) and low price (Gigabit Ethernet - GbE), as the bandwidth and latency gaps between Ethernet and Infiniband have been significant. However, the emergence of 10 Gigabit Ethernet (10GbE) closes this performance gap and provides a compelling alternative. 10GbE leverages the highest volume and lowest cost networking technology, and is starting to ramp in volume and drive down the cost curve. Already cheaper than Infiniband in absolute terms and more cost-effective than GbE in price/performance terms, 10 GbE offers other attractive benefits, including:

- Network flexibility. Because 10 GbE is well on its way to becoming the network of choice in the data center, it can be used not only for inter-processor communications, but also for general networking and storage. In an HPC network, 10 GbE can be used for small and large clusters alike. As a result, network administrators have flexibility to dynamically configure compute clusters according to the needs of a particular application, without being constrained by pre-configured Infiniband and GbE compute nodes.
- Backwards compatibility. 10 GbE is fully backwards compatible with GbE, so it can fully leverage the existing networking infrastructure, while allowing compute nodes and new switches to provide much higher performance.
- Easy to deploy and manage. 10 GbE is not only compatible with existing network infrastructure, it is also fully compatible with existing applications, so it can be deployed seamlessly with minimal effort and cost.

A key driver in the transition to 10 GbE is the significant increase in capabilities of the latest generation of servers. These new servers, powered by the latest processors with 8 cores each or more, provide a dramatic increase in compute capabilities. This in turn drives the need for increased network I/O capabilities to keep pace with improvements in

application processing. 10 GbE not only provides a much higher bandwidth pipe, but has also proven to be extremely efficient at network I/O delivering high performance while minimizing resource utilization. This performance capability, combined with an aggressive cost curve that continues to push pricing down illustrates why 10 GbE will become a standard feature on server motherboards in the coming years. The bottom line is GbE is no longer sufficient nor the most cost-effective solution for HPC clustering.

This paper illustrates how 10 GbE closes the Infiniband performance gap and offers a significant price/performance value proposition over GbE.

THE CASE FOR 10 GIGABIT ETHERNET IN HPC APPLICATIONS

When compared to legacy GbE, Infiniband can provide 10x performance benefits in terms of both bandwidth and latency. However, these performance benefits come with increases in both capital and operational costs. Infiniband equipment is procured at a 5x cost per port over GbE, while the additional cost of learning and installing different networking hardware and modifying the IP-based applications further increases the IT budget. To manage costs, Infiniband networks are used judiciously and only where the highest performance HPC clusters are required, leaving little or no headroom in the infrastructure. As a result, Infiniband-based clusters are a relatively fixed resource that cannot be easily expanded, so configurations are relatively inflexible.

Where low cost is needed, GbE cluster networks are used. These clusters typically use the GbE network port available on the server and a cost-effective Ethernet switch. Installing, configuring, and supporting these cluster networks is relatively easy and low-cost since it leverages general networking expertise already available in most organizations, and does not involve installing and configuring special transport protocols. In addition, these networks provide the most flexibility, as cluster size can be easily adapted to the size required. The trade-off using GbE is performance – with one tenth the bandwidth and 10 times the latency of Infiniband – this is a significant compromise.

The emergence of 10 GbE has provided a viable alternative for HPC cluster networks, providing users with significant advantages, including:

- Ease of installation and management
- Raw high performance – high bandwidth/
low latency
- Exceptional price/performance
- Scalability
- Low power consumption
- Performance/watt

Ease of Installation and Management

Based on the ubiquitous Ethernet standard, 10 GbE offers the same interoperability and compatibility as GbE offers, and leverages the same management techniques, so IT personnel can readily upgrade their physical networks to the higher speed version with little or no impact.

Another aspect of ease of installation is cabling choice. 10 GbE utilizes a variety of connector and cabling options. A popular option today is SFP+, which uses a small-form factor module to connect to a cable. The most popular module types are: optical transceiver that connects to fiber cabling; direct attach copper, which is an active cable with an SFP+ connector that plugs directly into the SFP+ cage; 1000BASE-T SFP module, which includes a physical layer transceiver that is compatible with the 1000BASE-T standard, and adapts RJ45 connectors to a standard SFP+ cage.

In addition to SFP+, 10 GbE server adapters are also available with native 10GBASE-T, which leverages existing Ethernet cabling and RJ45 connectors. 10GBASE-T provides backwards compatibility and automatic speed negotiation, which enables HPC cluster nodes to utilize existing GbE switching infrastructure or upgrade to 10 GbE switches. This ability to upgrade servers and server adapters independently from network infrastructure provides important flexibility in managing purchase costs and upgrade cycles.

High Performance, High Value, and Low Latency

Table 1 compares the performance and cost between GbE and 10 GbE alternatives of HPC cluster networks being discussed. As the table shows, 10 GbE improves GbE bandwidth by 10x. What is less obvious is the dramatic 5x reduction in latency 10GbE offers, which becomes critical in high traffic networks.

Additionally, 10GbE enables scalable I/O for high bandwidth applications with modest CPU utilization that cannot be addressed by GbE. For example, Solarflare testing has demonstrated that application level TCP/IP throughput in excess of 120Gbps is possible on a quad socket Nehalem-EX platform while consuming only 25% CPU and 38Gbps is easily achieved on single socket (Westmere) servers at approximately 20% utilization. This level of network bandwidth cannot be attained by scaling multiple GbE ports, and unlocks the I/O potential of multi-core servers, enabling large cluster scaling for many HPC codes without the use of proprietary interconnects.

Most importantly, on a bandwidth-adjusted basis 10 GbE is less expensive to acquire than GbE.

	1Gb Ethernet	10Gb Ethernet
Bandwidth	1 Gbps	10 Gbps
Latency (1/2 RTT)	20 μ sec	4 μ sec
Efficiency (Gbps/%CPU)	0.93	10.00
Price Per Gbps/Port	\$160	\$66

Table 1: HPC Interconnect Comparison: GbE vs. 10GbE

Compared to Infiniband, 10 GbE performance and cost compare favorably. Additionally, 10 GbE leverages GbE installation cost, which eliminates both special training and application modification, thus significantly lowering OpEx in addition to CapEx.

Greater Power Efficiency, Performance/Watt

Table 1 also compares the power efficiency between GbE and 10 GbE, reflecting that the 10 GbE products available today are extremely power efficient. For example, a single port of 10 GbE can deliver 5x the bandwidth at the same power consumption level as an integrated dual-port GbE LAN on motherboard (LOM) device. Furthermore, the GbE LOM can be disabled and replaced with a 10GbE server adapter which only consumes 2x the power and is far more power efficient than scaling bandwidth using multiple GbE ports. This superior power efficiency enables far greater performance and cluster scaling, as power and cooling limitations typically become an issue.

10 GIGABIT ETHERNET HPC MIGRATION

These technology and business reasons are supported by market analyses that suggest a strong migration of HPC deployments to 10 GbE in the next three years, from today's dominant Infiniband and GbE deployments. As a reference point, today GbE and Infiniband each hold approximately 40% share of the HPC cluster interconnect market, and will ship approximately 1.3 million ports into this market in 2010, according to IDC. Although not yet a significant factor in HPC clustering, in 2010 overall server-based port shipments of 10 GbE are expected to exceed 2.7 million ports, according to Dell'Oro. Looking ahead just a few years to 2013, Infiniband server port shipments are expected to roughly double, while overall 10 GbE server port shipments are forecast to grow more than 8x to over 23 million ports. Analogous to the 1GbE transition, 10 GbE is on the leading edge of a steep price curve that will make it a compelling, cost-effective technology. Since 10 GbE provides substantial increases in performance (latency and bandwidth), ease of installation and management, and full application compatibility, the trade off between performance and cost is no longer necessary – 10 GbE provides both.

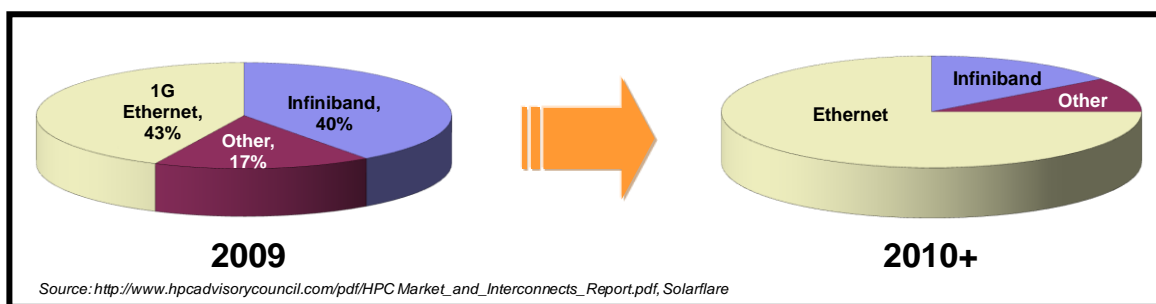


Table 2: HPC Interconnect market trends

SOLARFLARE 10 GIGABIT ETHERNET TECHNOLOGY LEADERSHIP

While 10 GbE provides significant performance and cost advantages over the alternatives, like all technologies some implementations are better than others. Solarflare server adapters deliver the industry's lowest power, lowest latency, lowest CPU utilization, and highest application performance, along with industry leading scalability.

Benchmark results show Solarflare's technology beats all other competitors in both bandwidth and latency. For example, Table 3 shows results of performance testing using MPI 7 (HP MPI) cluster stack, running over 64-bit RHEL 5.5 on a two-way quad-core 2.4GHz Westmere (E5620) with 6x 1Gbyte DIMMs. This testing compares the performance of Solarflare's SFN5122F server adapter with that of adapters from popular HPC competitors. The results show that Solarflare's

SFN5122F using OpenOnload application acceleration software outperforms all other competitors. Further, when compared to server adapters using RDMA, the SFN5122F delivers 30% lower latency than its nearest competitor. These results show that with Solarflare, the highest performance can be achieved without compromise – no need to modify applications, upgrade the network, use proprietary or dual-ended protocols, or other techniques that require modifications or specialization.

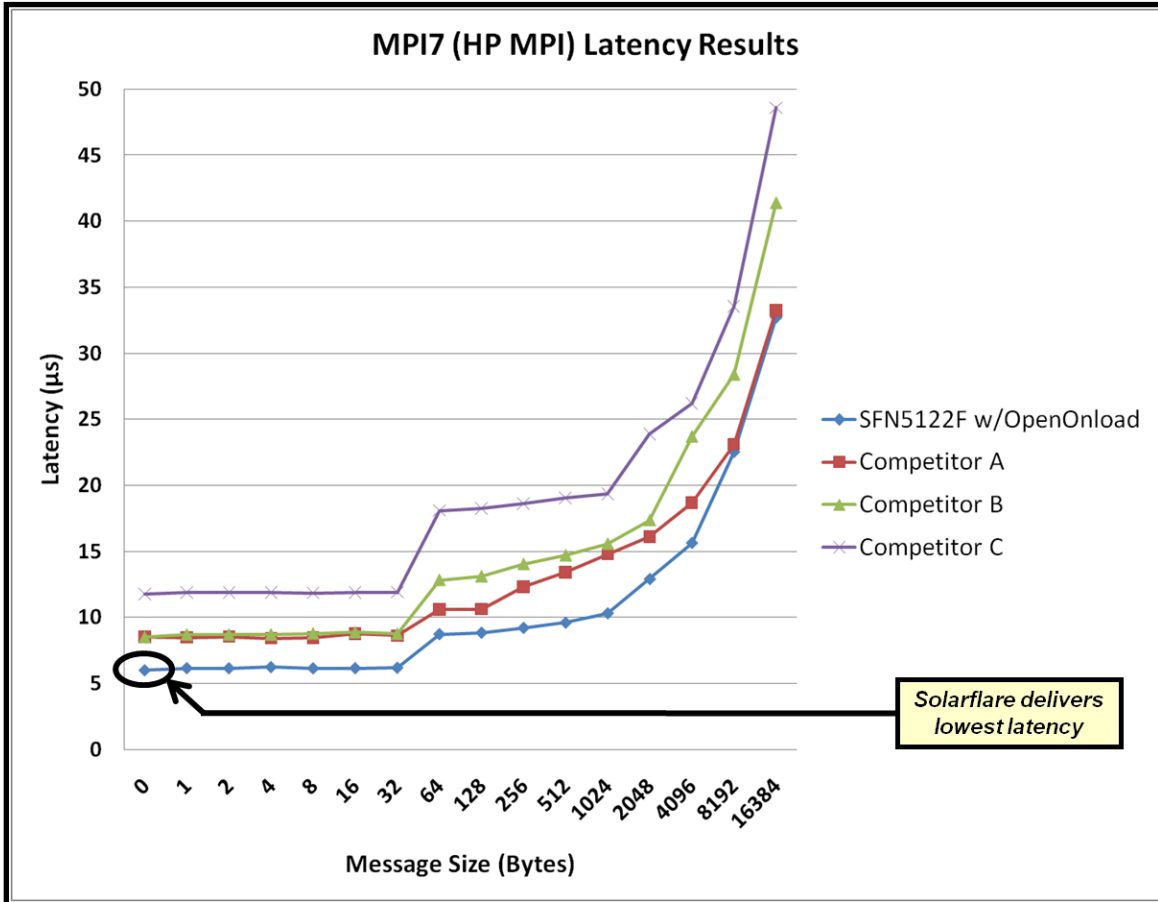


Table 3: HP MPI Send/Receive latency results

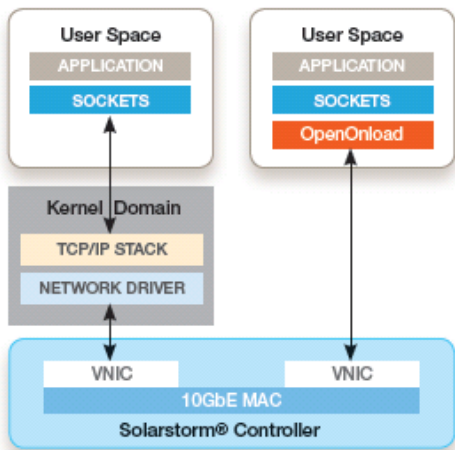
SUMMARY

Solarflare 10 GbE server adapters provide exceptional price/performance for HPC networks. Delivering latency to within 2µs of Infiniband and 40Gbps aggregate bandwidth for a dual-port server adapter, 10 GbE is more cost-effective than GbE, and easily replaces or scales out GbE installations.

**OPENONLOAD APPLICATION
ACCELERATION MIDDLEWARE FURTHER
LOWERS LATENCY**

Solarflare server adapters deliver the lowest kernel latency performance on the market, and OpenOnload significantly improves that latency advantage. When combined with Solarflare’s OpenOnload application accelerator middleware, Solarflare 10 GbE adapters can deliver sub-4 microsecond TCP/UDP application to application latencies, while supporting message rates in the millions, reducing latency jitter, and bringing a greater level of predictability to message processing latency.

OpenOnload is binary compatible with the BSD sockets API, requires no modification of the end user’s application, and because it is completely compatible with TCP/IP and Ethernet, requires no new wire protocols nor upgrades to the network.



SOLARFLARE’S FAMILY OF HPC SOLUTIONS

Solarflare offers single- and dual-port 10 GbE server adapters that deliver high bandwidth, industry leading latency and power, with stateless offloads that minimize CPU utilization. The Solarflare family supports both SFP+ and 10GBASE-T media. The SFP+ adapter supports optical modules or direct attach copper twin-ax cables, while the 10GBASE-T supports Category 6A, 6, 5E cables which are compatible with existing data center infrastructures for distances up to 100 meters.

Solarflare’s two families of server adapters meet a broad range of HPC networking needs. Enterprise server adapters are targeted at applications demanding the lowest latency, and most scalable virtualization. Midrange server adapters offer an exceptional 10GbE value.

Part number	SFN4112F	SFN5152F	SFN5162F	SFN5122F	SFN5151T	SFN5161T	SFN5121T
# OF PORTS	1	1	2	2	1	2	2
PCIe x8	2.5GT/s (Gen1.1)	5.0GT/s (Gen2)	5.0GT/s (Gen2)	5.0GT/s (Gen2)	5.0GT/s (Gen2)	5.0GT/s (Gen2)	5.0GT/s (Gen2)
CONNECTOR	SFP+	SFP+	SFP+	SFP+	RJ45	RJ45	RJ45
CABLING	DA, MMF	DA, MMF	DA, MMF	DA, MMF	UTP, STP	UTP, STP	UTP, STP
POWER (TYP)	4.5W	4.9W	4.9W	4.9W	< 13W	< 13W	< 13W
DIRECT GUEST ACCESS	✓			✓			✓
SR-IOV	✓			✓			✓
OPENONLOAD	✓			✓			✓
# OF VNICS	4096	128	128	2048	128	128	2048

ABOUT SOLARFLARE COMMUNICATIONS, INC.

Solarflare Communications is the leading provider of 10 Gigabit Ethernet (10GbE) silicon and server adapters. Solarflare’s robust and power-efficient solutions are cost effective and easy to deploy. Ready for primetime, Solarflare 10GbE products and OpenOnload™ make possible next-generation applications such as low-latency networking for market data applications, cloud computing, server virtualization, and network convergence.

For more information on Solarflare, please visit www.solarflare.com.

WORLDWIDE OFFICES

USA:	EMEA:
Solarflare Communications 9501 Jeronimo Road, Suite 250 Irvine, CA 92618, USA Phone: +1 949.581.6830 x2050 Fax: +1 949.581.4695 Email: sales@solarflare.com	Solarflare Communications Development Office Westbrook Centre, Building 2, Milton Road Cambridge UK CB4 1YG Phone: +44 (0)1223.518040 x5530 Fax: +44 (0)1223.464225

www.solarflare.com/industry/partner_buy.php